

Day at a Glance

Los Alamos National Labs: IBM Big Data Briefing		
Los Alamos Research Park, Room 203A		
Thursday, August 15, 2013		
<b>8:00 - 9:00</b>	<b>Registration &amp; Check-In</b>	
<b>9:00 - 9:15</b>	<b>Welcome Opening Remarks</b>	<b>David Wiseman</b>
<b>9:15 - 10:45</b>	<b>Main Tent Speaker - IBM Big Data Overview</b>	<b>Tim Paydos</b>
<b>10:45 - 11:00</b>	<b>Break</b>	
	<b>Track 1: Data Platforms &amp; Infrastructure</b>	<b>Track 2: Data Veracity, Analytics, and Reporting</b>
<b>11:30</b>	<u>Data Management &amp; Monitoring &amp; Movement:</u> C. Maestas -Platform Computing -GPFS (stand-alone) -Data Movement	<u>Dealing with Uncertain Data at Scale-</u> Paul Giangarra
<b>12:30</b>	<b>Invitation Only Working Lunch Speaker/ Facilitator:</b> <b>Paul Giangarra</b>	
<b>12:30 - 1:30</b>	<b>"Leveraging Cloud Computing for Research"</b>	
<b>1:30</b>	<u>BigData Analytics in Motion and at Rest</u> - Alain Biem/John Rollins	<u>Advanced Decision Support Combining Structured and Unstructured Analytics</u> - Paul Giangarra
<b>3:00</b>	<b>Break</b>	
<b>3:00 - 3:15</b>	<b>Break</b>	
<b>3:15</b>	<u>Integration &amp; Case Studies</u> - Biem/Maestas -Platform & BigInsights & Streams working together -Sensor based work & commercial use cases	<u>Predictive Analytics &amp; Reporting</u> - Deepak Nelli
<b>4:45</b>	<b>Wrap Up</b>	
<b>4:45 - 5:30</b>	<b>David Wiseman</b>	

## Track 1:

### **Data Platforms and Infrastructure**

*This section introduces the audience to inherent problems related to Big Data and provide an overview of the IBM Big Data Platform. The IBM Big Data platform is a stack of technologies optimized for Big Data computation including analytics development and deployment, workload optimization, scheduling, and visualization. The IBM Big Data Platform is designed to empower the analyst, the scientist, and the developer to tackle various aspects of Big Data.*

#### ***Data Management and Monitoring***

11:30  
to  
12:30

In this first session, we will cover some of the underlying technologies that are used to manage data within the resource pools we are scheduling and monitoring user workloads. The focus talks about 1) Platform Symphony which is a workload system that enables highly efficient job scheduling of applications that operate on data at rest, 2) GPFS technologies in features with regards to Flash integration, File Placement Optimization, and Integration Lifecycle Management examples, and 3) Platform Cluster Manager, which features xCAT and enables Big Data environments to be deployed and monitored.

#### ***Big Data Analytics in Motion and at Rest***

1:30  
to  
3:00

In the second part of the section, we introduce the IBM Big Data Platform as a stack of technologies that deal with various aspects of Big Data. We focus on 3 key technologies on that platform: (a) the IBM InfoSphere BigInsight, which is the IBM MapReduce solution for unstructured and structured Big Data at rest; (b) the IBM InfoSphere Streams, an innovative and robust middleware designed to facilitate development and deployment of analytics for Big Data in motion, and (c) the IBM PureData System for Analytics, an integrated, massively parallel data warehousing platform providing powerful in-database analytics for Big Data.

#### ***Integration and Case Studies***

3:15  
to  
4:45

This session complements the previous session by highlighting proven solutions that use the IBM Big Data platform. We will discuss and explain various projects and customer engagements that rely on IBM BigInsight and IBM Streams technologies. Those projects span various industries and domains including healthcare, physical sciences, radioastronomy, network analysis, social media, energy, or traffic management. We also discuss how Platform Symphony, GPFS and Platform Cluster Manager can help strengthening these use cases.

## Track 2:

### Data Veracity, Analytics, and Reporting

*The journey from raw data to information, understanding, and insight.*

#### **Dealing with Uncertain Data at Scale**

11:30

to

12:30

The first session will review key principles of big data and the discuss the concepts of Data Veracity and dealing with Uncertain Data at Scale. This session will also include a discussion of patterns, technologies (and products), and best practices that can be applied to business problems. When these are applied the "uncertain data" will be collected, cleansed, and ready for various types of analytics. Supporting technologies include data quality with *InfoSphere Quality Stage*, normalization with *WebSphere Message Broker*, data cleansing with *InfoSphere Streams*.

#### **Advanced Decision Support Combining Structured and Unstructured Analytics**

1:30

to

3:00

The second session will discuss advanced decision support systems. This section will start with detailed discussion of the A/FDIR (Anomaly/Failure Detection, Isolation, and Recovery) This includes a discussion of "record and retrieve", various data cleansing, data normalization, data mediation, complex event processing, structured decision logic (rules based and state machine based), and the incorporation of unstructured decision support (based on IBM's Watson technologies). It will also include the use of predictive and/or prescriptive analytics to help rank solution options, and to analyze the consequences of each solution option. Technologies discussed include *Watson*, *IBM Content Analytics*, *WebSphere Operational Decision Manager*

#### **Predictive Analytics and Reporting**

3:15

to

4:45

The third and final session will discuss IBM's Descriptive and Predictive analytics capabilities. This includes standard as well as advanced visualization, based on *Cognos*, and predictive analytics based on *SPSS*.

## Invitation Only Working Lunch

### **Leveraging Cloud Computing for Research**

12:30

to

1:30

*In today's scientific world where projects are dynamic, computers provide the resources for intensive data analysis as well as standard IT functions (such as word processing, content management, etc.). Cloud Computing becomes a means to share expensive and powerful IT resources across many projects. This includes both the tools and the physical resources needed by those tools. For example, when a compute intensive model needs to be run, the modelling tool can be selected for use, possibly with some visualization and reporting tools. Then the compute resources can be specified and the "cloud" will dynamically provision the tools and physical resources for the duration of time they are needed. When they are done with they can be returned to the cloud, much like a reference library book borrowed and then returned to the library, to be utilized another day by someone else or even the same research project.*

## **IBM Presenters**



**Paul Giangarra**  
**IBM Distinguished Engineer**

Paul Giangarra is a Distinguished Engineer and e-Government Solutions Architect in IBM Federal's Office of the CTO. In this role, Paul works with nearly all branches of Federal government including NASA, the FAA, the Department of Homeland Security, the Department of Defense, Intelligence Agencies, and other civilian agencies helping them understand, identify, and solve critical and complex IT problems. He also works closely with FFRDCs like JPL and MITRE. Internally at IBM Paul works across IBM with Research, Development, Hardware, Services, and Corporate CIO teams on a variety of projects. He is an expert on topics such as SOA, ESB, Cloud Computing, Information Management, Governance, Complex System Integration, Java for the Enterprise, Real-Time, Event Driven Architecture, Decision Support Systems, Analytics (especially DeepQA) and Enterprise Architecture. Paul is a member of the IBM Software Group Architecture Board Steering Committee. Paul is a member of the IBM Academy of Technology Leadership Team. Paul is regularly in demand by organizations such as IEEE, AFCEA, as well as government agencies and Federal integrators and FFRDCs for panels and as a speaker.

[www.linkedin.com/pub/paul-giangarra/9/868/7](http://www.linkedin.com/pub/paul-giangarra/9/868/7)



**John B. Rollins, Ph.D., P.E.**  
**Chief Data Miner and Technical Manager**

John is the Chief Data Miner and Technical Manager of the Advanced Analytics POC Team, IBM Netezza Analytics Solutions, IBM Software Group. His background is in the fields of engineering, econometrics, and advanced analytics in many industries. He holds seven patents, and he has authored a best-selling engineering textbook, many peer-reviewed technical papers, and two IBM Redbooks. He holds two doctoral degrees in economics and petroleum engineering from Texas A&M University and is a registered professional engineer in Texas.



**Tim Paydos**  
**Director, Worldwide Government Information Agenda Team**

As Industry Leader, IBM Business Analytics & Optimization and Information Agenda, Tim Paydos is responsible for IBM's strategy in support of government agencies at all levels who seek to leverage the value of their information assets to address their most critical information intensive business problems. Previously, Tim was Director, IBM InfoSphere, where he was responsible for IBM's strategy in support of government agencies and commercial leaders who seek to establish, manage, govern and leverage Trusted Information. He is a seasoned solutions and government professional with over 15 years in the field. Through his career Tim has worked directly with dozens of clients across the Social Services, Tax and Revenue, Justice, Intelligence and Defense Space. Experiences outside IBM include Customer Business Development and Operations for Procter & Gamble, Vice President of Strategy at Syncra Systems. Tim holds a BA from Harvard University, lives in Connecticut with his two daughters, and for relaxation serves as a nationally certified fire/rescue instructor and Captain in the Simsbury Fire Department.

[www.linkedin.com/pub/tim-paydos/0/197/857](http://www.linkedin.com/pub/tim-paydos/0/197/857)



**Alain Biem, PhD**  
**Big Data/Streaming Analytics - R&D Lead**  
**Thomas J. Watson Research Center**

Alain Biem is a Senior Research Scientist and Project Lead at IBM Research in New York, where he has contributed or led various research and development projects in the area of Big Data, predictive analytics, time series processing, distributed processing, information management, and business modelling. Alain has been involved in customer-facing projects spanning diverse industries and domains, including human speech or human script recognition by machine, environmental data processing (radio-astronomy, seismic, high-energy data), traffic monitoring, healthcare, and cyber-security.

Alain has received numerous awards for his contribution to various leading IBM products including the Cell Broadband Engine (the core chip in the Playstation 2), the innovative IBM Transnote computer, the IBM Business Component Modeling (CBM), online healthcare analytics, and the IBM InfoSphere Streams platform. He is doing research in automated analytics Design and Deployment and is the technical lead and architect of the Time Series analysis toolkit for IBM Infosphere Streams Big Data product.

Prior to joining IBM in 2000, Alain was a research scientist at NTT/ATR, Kyoto, Japan working in the area of English and Japanese speech recognition. Alain received his PhD with Summa Cum Laude honors in computer science (machine learning) from Paris 6, France. His MS in EE from Telecom Grande ecole in France (mathematique superieure and speciales). Alain has numerous publications in leading signal processing and machine learning conferences and journals, is a committee member in various IEEE or ACM conferences, and has directed PhD students in Europe, Canada, and US.

[www.linkedin.com/in/aebiem](http://www.linkedin.com/in/aebiem)



**Deepak Nelli**  
**Technical Business Unit Executive, Public Sector**  
**IBM Business Analytics**

Deepak Nelli is the Technical Business Unit Executive for IBM Business Analytics. For more than 15 years Deepak has worked extensively with public sector entities across the United States providing them with successful strategies and solutions to meet their business analytics needs. At IBM, Deepak is responsible for managing the national public sector technical sales team and providing direction for IBM's Business Analytics products by communicating the public sector strategy and vision.



**Chris Maestas**  
**Technical Sales Specialist – HPC, Cloud, and Grid Solutions**  
**IBM Systems and Technology Group**

Chris is a Field Technical Sales Specialist in the GPFS and Platform Computing product system engineering group. For more than 15 years, Chris has worked with entities that span customers in Federal, State, Local, Education and Commercial spaces addressing their HPC challenges. At IBM, Chris is responsible for helping customers meet their objectives in data and information management. Chris has a B.S. in Computer Science from the University of New Mexico.

[www.linkedin.com/in/cdmaestas](http://www.linkedin.com/in/cdmaestas)